Cell Host & Microbe
Commentary

Computational Approaches to Influenza Surveillance: Beyond Timeliness

Elaine O. Nsoesie[1,2,*] and John S. Brownstein[1,2,3,*]
[1]Children's Hospital Informatics Program, Boston Children's Hospital, Boston, MA 02115, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC H3A 1A2, Canada
*Correspondence: onelaine@vt.edu (E.O.N.), john.brownstein@childrens.harvard.edu (J.S.B.)
http://dx.doi.org/10.1016/j.chom.2015.02.004

Several digital data sources and systems have been advanced for use in augmenting traditional influenza surveillance systems. Although timeliness is one of the main advantages of these tools, there are several other recognizable uses and potential impact of these systems on the public and global public health.

Introduction

The field of digital disease surveillance commenced with the analysis of online news media in the mid-1990s and has evolved over the years to include a variety of text-based and non-text-based sources (Hartley et al., 2013; Salathé et al., 2012). Digital disease surveillance systems traditionally gather, process, and disseminate digital data on public health issues. Early examples include ProMED (Program for Monitoring Emerging Diseases), a system introduced in 1993 (Madoff and Woodall, 2005) that currently uses mailing lists and listserv subscriptions to assemble and disseminate information on disease outbreaks (including plant and animal diseases), infectious disease expert commentary from field clinicians, public health workers, and moderated news reports. Over the last 20 years, several systems using diverse data sources, with varying geographical coverage (ranging from the local to the international) and disease focus have been developed. These systems are typically built to enhance traditional indicator-based surveillance systems with the potential to aid in medical decision making, improve assessment of population response toward disease control (e.g., vaccination sentiments), understand disease spread relative to population density and movement, and aid in the early detection of disease events, including those emerging from remote regions (Brownstein et al., 2009; Hartley et al., 2013; Salathé et al., 2012). Discussions on the use of digital surveillance systems, challenges and limitations, and future research that could aid to improve the usage of these systems have been published (Hartley et al., 2013; Milinovich et al., 2014; Morse, 2012; Salathé et al., 2012).

Several of the existing digital disease surveillance systems have been used for monitoring influenza and influenza-like illness, and new systems are frequently introduced. The extensive interest in applying computational approaches to influenza surveillance has led to the exploration of various online data sources, digital technologies, and computational and data mining techniques. However, it is worth noting that the majority of these systems probably capture "influenza-like" illness, which may be driven by a range of respiratory pathogens.

In addition to timeliness, which is typically advanced as a main improvement of these tools over traditional public health surveillance, there are several advantages especially for surveillance in data-poor regions. Here, we summarize established approaches, discuss recent advances, and examine the known and potential utility of these tools.

Systems and Data Sources for Influenza Surveillance

Computational approaches for influenza surveillance can be broadly categorized as *active* and *passive*. Active surveillance is defined here as the targeted collection of information from the population, such as crowd-sourcing using cell phone apps and participatory approaches. In contrast, passive surveillance can be described as the extraction of existing data from sources such as specific web pages using machine learning techniques (e.g., crawling and scraping).

Participatory Surveillance Systems

The first participatory surveillance system for influenza, de Grote Griepmeting, was introduced in the Netherlands and Belgium in 2003. Since then, there have been several participatory surveillance systems for influenza developed for different countries (see Figure 1). In 2008, Influenzanet, a European-wide consortium for monitoring influenza-like illness using participatory surveillance systems was established (Paolotti et al., 2014). The Influenzanet network is composed of the United Kingdom, Sweden, Spain, Ireland, the Netherlands, Belgium, Denmark, Italy, Portugal, and France. The consortium aims "to rapidly identify public health emergencies, contribute to understanding global trends, inform data-driven forecast models to assess the impact on the population, optimize the allocation of resources, and help in devising mitigation and containment measures" (Paolotti et al., 2014). Although there are some differences in the participatory surveillance systems presented in Figure 1, these systems typically collect some background information at time of registration and send surveys to registered participants at regular intervals, usually weekly, to gather data on disease symptoms experienced during the previous week. The symptoms data are processed and presented using maps or other methods aimed at informing the public of influenza-like illness activity levels. Challenges to participatory surveillance include recruiting and maintaining participants, accuracy of self-reported data, developing a nationally representative sample, and specifically monitoring at-risk populations. Despite these

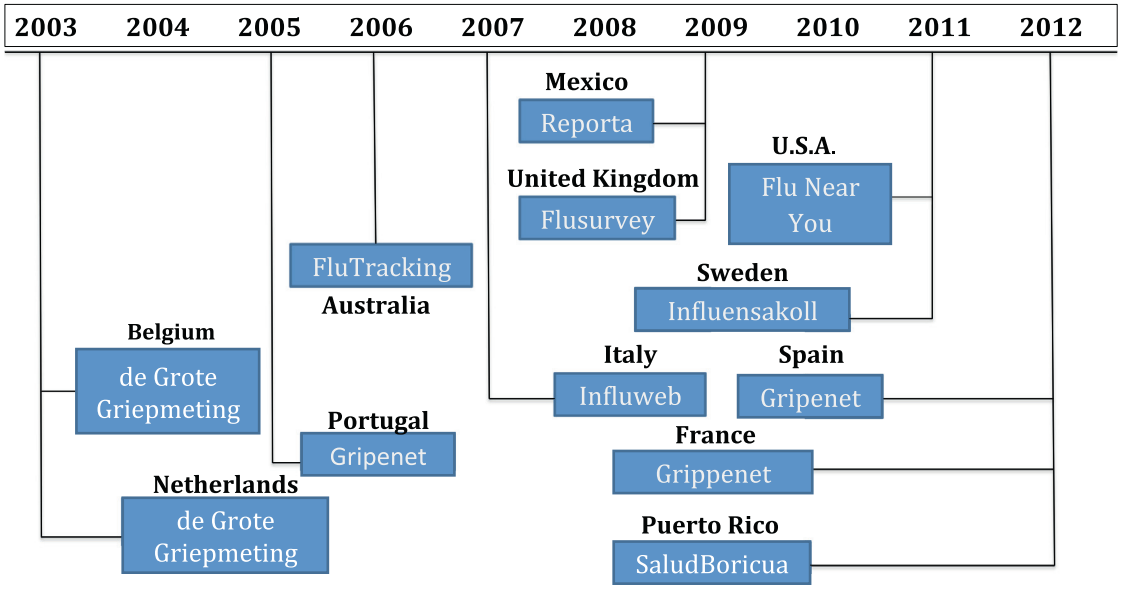

Figure 1. Participatory Surveillance Systems for Influenza-like Illness and Date of Launch

limitations, data from these systems have been shown to have similarities in trends and peak timing when compared to reports from practitioner-based surveillance systems (Paolotti et al., 2014). Furthermore, data from these systems have also been used to assess vaccination coverage and inform epidemiological models for influenza-like illness (Paolotti et al., 2014; Wójcik et al., 2014).

Internet News Data Systems

One of the earliest examples of an Internet news-based data system is the GPHIN (Global Public Health Intelligence Network) developed by the Public Health Agency of Canada. Other examples include HealthMap, MediSys, and BioCaster. Structured and unstructured information for the aforementioned systems are extracted from unofficial sources (e.g., online news sources, blogs, and social media) and official sources (e.g., ministry of health webpages and international public health organizations). Although the data collection process for these systems vary, they usually include data procurement from the Internet, processing using automated and semi-automated processes to detect trend and anomaly, assembling of information at a spatial and/or temporal scale, and dissemination to the public and public health practitioners (Brownstein et al., 2009; Hartley et al., 2013). Information extracted includes the disease name, affected species, location and date of outbreak, and case data, including counts on suspected and confirmed cases. Some systems are disease specific, while others are focused on extracting and gathering information on all communicable disease events. These systems also cover information at different geographical scales, from local to international. Due to the unstructured nature of some of the information retrieved, the data retrieval process can be challenging (Hartley et al., 2013; Salathé et al., 2012), and sometimes trained data "curators" manually correct misclassifications that are then applied through an iterative process to improve the machine learning algorithms used in data classification. Given the large amount of data collected by these systems, events of public health importance are often buried within many reports of less severe disease outbreaks. If detected in a timely manner, unusual events can be further investigated and public health risk effectively assessed.

Search Query Systems

In addition to the aforementioned systems, there have been several studies assessing the use of population web search records to estimate and predict influenza-like illness activity (Morse, 2012; Nsoesie et al., 2014). Initial studies published in 2008 and 2009, respectively, evaluated the use of web searches from Yahoo and Google for estimating influenza activity (Milinovich et al., 2014). Studies have also evaluated queries from clinician support tools and medical websites for monitoring trends in influenza-like illness. These studies have shown significant correlations between data from the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and clinician searches on medications, respiratory viruses (such as adenovirus, rhinovirus, coronavirus, etc.), and influenza-related terms. Additionally, search query data have also been used in the estimation of different measures of the influenza epidemic curve (e.g., intensity, peak time, and incidence) and as input into disease transmission models to predict influenza spread. Google Flu Trends (http://www.google.org/flutrends/us/), introduced in 2009, and HealthMap FluCast (http://healthmap.org/flucast/), initiated in 2014, are two digital surveillance systems that rely on search query data for influenza-like illness surveillance. Time series data representing searches of particular keywords are used to model and predict influenza-like illness reports from official sources such as the U.S. Centers for Disease Control and Prevention (CDC). Google Flu Trends uses a single data source, namely, searches submitted through the Google search engine, while HealthMap FluCast uses data from Google Trends and other sources (such as electronic health records

from AthenaHealth). Sometimes information produced using these sources varies from that presented by traditional surveillance systems (Salathé et al., 2012). There are several potential reasons for this, including changes in the underlying data-generating tools and population health-seeking behavior, as well as discrepancies between sick individuals and individuals searching for influenza-related terms. Errors in estimates and predictions from digital disease surveillance tools can be misleading to the public and can undermine the potential of and confidence in these systems. As a consequence, these systems need to be regularly updated and validated to minimize real-time over- or under-estimation of influenza-like illness. Once developed, however, the cost of running and maintaining these systems is relatively low and the data sources used in most instances are openly available.

Social Media Systems

Data from social networking sites have also been shown to have potential for influenza monitoring and prediction. Initial studies assessed the use of reports of influenza-like illness on Twitter (a micro-blogging site) to assess spread of influenza-like illness during the 2009 A (H1N1) influenza pandemic in the United States. There have been several studies focused on using illness reports on Twitter for seasonal influenza surveillance and forecasting. Content from Twitter mentioning influenza or influenza symptoms are extracted and analyzed to estimate disease spread both temporally and spatially. Two examples of systems that use social media data for monitoring influenza-like illness are Sickweather (http://www.sickweather.com) and FluCaster (http://ndssl.vbi.vt.edu/apps/flucaster/). Sickweather combines self-reported information with geolocated data from social networking sites such as Twitter to provide information on the spatial spread of influenza-like illness. FluCaster also uses crowd-sourcing and social media for influenza-like illness surveillance. However, FluCaster further utilizes a complex computational epidemiology model, which enables estimation of the probability of infection and assessment of the effectiveness of different intervention strategies. Sickweather and FluCaster were introduced in 2011 and 2013, respectively. Other systems such as HealthTweets.org aim to translate health research using social media into practice. Systems that use social media data for disease surveillance process large amounts of data to extract useful signals indicating disease activity. For some of these systems, there are disproportionate distributions of users across locations, age, and race/ethnicity that can lead to significant bias in the data sources. There are also concerns of data access, privacy, and data sharing when dealing with data from sources such as Twitter, Facebook, and Google. It is obvious that regulations are needed so that individuals' privacy is not violated and the data are used in an appropriate manner.

Other Data Sources

Recent studies on digital surveillance of influenza-like illness have evaluated the use of Wikipedia access logs for specific influenza-related articles, online reservation cancellations, and hospital traffic extracted from high-resolution satellite imagery. Studies using Wikipedia access logs have demonstrated statistically significant correlations between this data source and data from official sources. Significant correlations have also been recorded between trends in restaurant reservations and influenza-like illness activity for cities in Mexico and the U.S., suggesting that this data source could be useful for monitoring disease activity. A system developed to record reasons for reservation cancellations would function similarly to a participatory surveillance system. Lastly, other indicators such as hospital traffic extracted from high-resolution satellite imagery data can capture changes in population behavior due to an increase in the level of disease.

Usefulness and Potential of Influenza Surveillance Systems

Web-based disease information resources are used by major public health organizations (such as the WHO) (Chretien et al., 2008) and states and local communicable disease investigators for regular surveillance activities (M'ikanatha et al., 2006). Although usual attributes for assessing surveillance systems based on the effectiveness of response have been deemed inadequate (Paterson and Durrheim, 2014), there is some utility and potential impact of these systems on the public and global public health.

First, Internet-based data sources have been demonstrated to be valuable for detection, monitoring, and dissemination of information during recent influenza outbreaks (Salathé et al., 2013). While the timeliness of these systems might not have a significant observable impact during seasonal influenza epidemics, these systems are especially useful during epidemics resulting from novel influenza viruses. Digital disease detection systems have identified early reports of emerging influenza outbreaks. An example is the identification of a report of an unknown illness in Mexico, which was later determined to be caused by the A (H1N1) influenza virus (Brownstein et al., 2009). While news-based digital disease surveillance systems may capture early reports of disease outbreaks in rural regions, identifying these reports can be computationally intensive, costly, and challenging for real-time reporting. Internet-based systems can also aid in the dissemination of information on prevention during emerging influenza outbreaks and improve awareness of influenza and influenza-like illnesses through effective communication to the public (Wójcik et al., 2014)

Second, in addition to early detection of reports of disease, internet-based systems can be used for monitoring disease activity and extracting epidemiologic data on cases during an outbreak. For example, information aggregated through automated and manual processing from publicly available data sources during the H7N9 epidemic in China were shown to match official "line lists"—listing of infected persons with specific characteristics including demographic, clinical, and other epidemiologic data (Lau et al., 2014). Additionally, these systems have been recognized for encouraging government release of disease data and facilitating communication during emerging disease infections (Brownstein et al., 2009; Salathé et al., 2013). Geopolitical obstacles and communication barriers do not restrict the functioning of these systems.

Third, digital disease surveillance systems can aid in the understanding of spatial spread of influenza epidemics. By mapping reports of influenza and influenza-like illness, the public and

public health authorities can identify regions with the highest prevalence. In addition, data from systems that collect demographic information can be analyzed to better understand the impact of influenza on different demographic groups and compare spread across different communities for the implementation of targeted interventions. Furthermore, vaccination reports can be used to assess vaccine uptake and efficacy at different geographical scales (Wójcik et al., 2014).

Fourth, internet-based systems have been used to evaluate population health-seeking behavior and sentiments toward disease and disease control measures such as vaccination, which can be critical for the design and implementation of targeted control measures during influenza pandemics. These data can also enable a better understanding of changes in population behavior before, during, and after an outbreak. Novel data approaches such as high-resolution satellite imagery of disease-affected populations can provide a representation of how population behavior varies over time and can be used to assess response to specific intervention strategies such as social distancing. For example, high-resolution satellite imagery of hospital parking lots in Chile, Argentina, and Mexico were shown to be predictive of influenza activity at the national level (Butler et al., 2014). Further studies using targeted surveillance approaches can be useful in assessing changes in health-seeking behavior especially during major influenza outbreaks.

Fifth, disease-related data extracted from different sources could be compared and integrated to improve surveillance. The integration of data sources (e.g., the Internet and mobile phone technologies) can reduce gaps present in individual sources and systems. Data integration techniques using Bayesian ensemble and filtering methods have been shown to yield promising results both for influenza monitoring and prediction. The integration of diverse data sources or models based on a combination of different data types has the potential to improve estimates of influenza activity relative to a single system or data source. Ideally, validated systems could be integrated with existing healthcare surveillance infrastructure to "provide information about people who do not seek healthcare, data that is not otherwise available" (Wójcik et al., 2014).

Lastly, these systems and data sources have the potential to improve global public health by improving disease surveillance in data and resource poor regions. In such settings, data from supplementary sources (such as high-resolution satellite imagery and syndromic surveillance systems) could be integrated with data from traditional surveillance networks to identify unusual events and changes in morbidity and mortality trends, which could lead to the initiation of prompt investigation and response. The lack of a strong public health infrastructure implies that resource poor regions are less likely to have appropriate clinical resources for disease confirmation, which makes the case for using syndromic and similar surveillance techniques in such settings (Chretien et al., 2008).

Conclusions

There are several limitations to these systems, which we have previously highlighted. Specifically, they include (1) differentiating signal from noise; (2) significant biases due to differences in the representation of individuals from different locations, age, and race/ethnic backgrounds; (3) variations between information produced by internet-based systems and well-established official influenza surveillance systems; and (4) privacy and data access concerns. Additionally, it is still not yet well established how these systems could be structured to trigger alerts during influenza epidemics.

Despite these limitations and challenges, digital disease surveillance systems have the potential to aid in the monitoring of disease spread and communicating to public health practitioners and the public. If adopted by appropriate public health authorities, the data available through these systems can aid in timely detection and response, which is needed for disease control.

ACKNOWLEDGMENTS

E.O.N. is supported by funding from the National Institute of Environmental Health Sciences of the National Institutes of Health (Award Number K01ES025438). J.S.B. is supported by a research grant from the National Library of Medicine, the National Institutes of Health (5R01LM010812-05). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Brownstein, J.S., Freifeld, C.C., and Madoff, L.C. (2009). N. Engl. J. Med. *360*, 2153–2155, 2157.

Butler, P., Ramakrishnan, N., Nsoesie, E.O., and Brownstein, J.S. (2014). Computer *47*, 94–97.

Chretien, J.-P., Burkom, H.S., Sedyaningsih, E.R., Larasati, R.P., Lescano, A.G., Mundaca, C.C., Blazes, D.L., Munayco, C.V., Coberly, J.S., Ashar, R.J., and Lewis, S.H. (2008). PLoS Med. *5*, e72, http://dx.doi.org/10.1371/journal.pmed.0050072.

Hartley, D.M., Nelson, N.P., Arthur, R.R., Barboza, P., Collier, N., Lightfoot, N., Linge, J.P., van der Goot, E., Mawudeku, A., Madoff, L.C., et al. (2013). Clin. Microbiol. Infect. *19*, 1006–1013.

Lau, E.H.Y., Zheng, J., Tsang, T.K., Liao, Q., Lewis, B., Brownstein, J.S., Sanders, S., Wong, J.Y., Mekaru, S.R., Rivers, C., et al. (2014). BMC Med. *12*, 88, http://dx.doi.org/10.1186/1741-7015-12-88.

M'ikanatha, N.M., Rohn, D.D., Robertson, C., Tan, C.G., Holmes, J.H., Kunselman, A.R., Polachek, C., and Lautenbach, E. (2006). Biosecur. Bioterror. *4*, 293–300.

Madoff, L.C., and Woodall, J.P. (2005). Arch. Med. Res. *36*, 724–730.

Milinovich, G.J., Williams, G.M., Clements, A.C., and Hu, W. (2014). Lancet Infect. Dis. *14*, 160–168.

Morse, S.S. (2012). Biosecur. Bioterror. *10*, 6–16.

Nsoesie, E.O., Brownstein, J.S., Ramakrishnan, N., and Marathe, M.V. (2014). Influenza Other Respi. Viruses *8*, 309–316.

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., Koppeschaar, C., Rehn, M., Smallenburg, R., Turbelin, C., et al. (2014). Clin. Microbiol. Infect. *20*, 17–21.

Paterson, B.J., and Durrheim, D.N. (2014). Lancet Infect. Dis. *14*, 794.

Salathé, M., Bengtsson, L., Bodnar, T.J., Brewer, D.D., Brownstein, J.S., Buckee, C., Campbell, E.M., Cattuto, C., Khandelwal, S., Mabry, P.L., and Vespignani, A. (2012). PLoS Comput. Biol. *8*, e1002616, http://dx.doi.org/10.1371/journal.pcbi.1002616.

Salathé, M., Freifeld, C.C., Mekaru, S.R., Tomasulo, A.F., and Brownstein, J.S. (2013). N. Engl. J. Med. *369*, 401–404.

Wójcik, O.P., Brownstein, J.S., Chunara, R., and Johansson, M.A. (2014). Emerg. Themes Epidemiol. *11*, 7.